Univerza v Ljubljani
Fakulteta za računalništvo in informatiko

Jernej Lipovec

Uporaba strojnega učenja pri napovedovanju cen kart v igri Magic

diplomsko delo

univerzitetni študijski program računalništvo in informatika

Mentor: akad. prof. dr. Ivan Bratko

Ljubljana 2016

Besedilo je oblikovano z urejevalnikom besedil LaTeX na portalu ShareLatex.com.

Fakulteta za računalništvo in informatiko izdaja naslednjo nalogo:

Tematika naloge:

Magic: The Gathering je najbolj priljubljena zbirateljska igra s kartami, ki zabava igralce že več kot 20 let. Nekatere karte imajo zaradi redkosti in popularnost visoko vrednost, ki pa je do določene mere predvidljiva. Na ceno vpliva več dejavnikov, kot so igranost karte na turnirjih, bližina izida naslednje izdaje ter bližina velikih turnirjev. V diplomski nalogi analizirajte gibanje cen kart v izbranem časovnem obdobju ter s pomočjo metod strojnega učenja ustvarite sistem za napovedovanje gibanja cen kart glede na zunanje parametre. Identificirajte pomembne parametre napovedovanja cen in izdelajte program za pridobivanje teh podatkov. Pri diplomski nalogi uporabite splošno dostopna programska orodja za strojno učenje, kot je sistem WEKA.

Title of thesis:

Predicting the prices of cards in the game Magic with machine learning

Specification:

Magic: The Gathering is the most popular trading card game that has been played for over 20 years. Due to their rarity and popularity, some cards have a high value, which is predictable to some extent. The price of a card is affected by several factors, such as the use of the card in tournaments and the proximity of the next edition or important tournaments. In this project, analyse the trends of card prices in a chosen time period and employ machine learning methods to create a system for automatically predicting the trends in card prices depending on external parameters. Identify parameters important for price prediction and develop a program for collecting this data. In the project, apply generally available software tools for machine learning, such as WEKA.

Izjava o avtorstvu diplomskega dela

Spodaj podpisani Jernej Lipovec sem avtor diplomskega dela z naslovom:

Uporaba strojnega učenja pri napovedovanju cen kart v igri Magic

S svojim podpisom zagotavljam, da:

- sem diplomsko delo izdelal samostojno pod mentorstvom akad. prof. dr. Ivana Bratka,
- so elektronska oblika diplomskega dela, naslov (slov., angl.), povzetek (slov., angl.) ter ključne besede (slov., angl.) identični s tiskano obliko diplomskega dela,
- soglašam z javno objavo elektronske oblike diplomskega dela na svetovnem spletu preko univerzitetnega spletnega arhiva.

V Ljubljani, dne 16. marca 2016 Podpis avtorja:

Zahvaljujem se svoji punci Manci za vse nasvete o akademskem pisanju in podporo, ki mi jo je nudila pri pisanju. Zahvalil bi se tudi vsem kolegom, ki so prispevali k strokovni natančnosti naloge ter ožji družini za podporo med pisanjem. Velika zahvala gre tudi mentorju akad. prof. dr. Ivan Bratku za strokovno podporo in nasvete.

Kazalo

Seznam uporabljenih kratic

kratica	angleško	slovensko
TCG	Trading Card Game	Zbirateljska igra s kartami
MTG	Magic: The Gathering	Magic: The Gathering
SVM	Support Vector Machines	Metoda podpornih vektorjev
API	Application program interface	Vmesnik za programiranje aplikacije
MKM	Magic Card Market	Magic Card Market
DB	Database	Podatkovna baza
SMO	Sequential Minimal Optimization	Sekvenčna minimalna optimizacija

Povzetek

V tem diplomskem delu smo preučevali trende gibanja cen pri kartah igre Magic: The Gathering in pri tem uporabili najbolj primerne metode strojnega učenja. Cilj je bil izdelati napovedni model za cene kart. Naša naloga je bila identificiranje pomembnih virov, pridobivanje potrebnih podatkov, njihova pretvorba v računalniku razumljivo obliko ter izbira primernega algoritma. Model, ki smo ga ustvarili, se je izkazal za zanesljivega s 61% točnostjo napovedi gibanja cene pri zelo redkih kartah, medtem ko smo pri redkih kartah dosegli le 52% točnost, kar ni preseglo niti privzete točnosti. Pri nalogi smo uporabili metodo podpornih vektorjev ter si pomagali z orodjem Weka. S podatki, ki smo jih pridobili, smo naredili še nekaj poizkusov in tako poiskali nekaj novih odvisnosti med podatki, ki jih prej nismo poznali.

Ključne besede: strojno učenje, predvidevanje cen, MTG, prosti trg, ponudba in povpraševanje, metoda podpornih vektorjev, Weka, podatkovno rudarjenje.

Abstract

This thesis is a study of Magic: The Gathering card price fluctuations using the most appropriate machine learning methods. The goal was to construct a predictive model for card prices. This required us to identify crucial attributes, gather necessary data, convert it to a machine-readable format and select a suitable learning algorithm for the task. The resulting model was effective, attaining a 61 % price trend accuracy with mythic rare cards, while it was less successful with rare cards with only 52% accuracy, which failed to beat default accuracy. Support vector machines algorithms and the machine learning toolbox Weka were used to achieve these results, which were applied in further experiments that led to the discovery of previously unknown data dependencies.

Keywords: machine learning, price prediction, MTG, free market, supply and demand, support vector machines, Weka, data mining.

Poglavje 1

Uvod

Strojno učenje postaja iz leta v leto bolj popularna veda, njena uporabnost pa se kaže na več področjih, kot so priporočilni sistemi, avtomobili s samodejno vožnjo, biomedicina itn. Eno od takih področij je tudi predvidevanje nihanja cen raznovrstnih dobrin. Čeprav veda obstaja že več desetletij, je postala izjemno popularna šele v zadnjem desetletju, tudi s porastom interneta ter velikih količin podatkov, ki jih shranjujemo na naših računalnikih. Strojno učenje je trenutno vsekakor ena od najhitreje razvijajočih se ved v računalništvu, saj po njej posega vse več velikih podjetjih, na primer Google in AirBnB, povpraševanje po veščinah strojnega učenja pa že sedaj presega ponudbo.

V tej nalogi se bomo podali v svet kart Magic: The Gathering, kjer bomo odkrivali zakonitosti v podatkih s pomočjo sodobnih tehnik strojnega učenja. S porastom trgovanja kart na internetu so nam na voljo velike količine podatkov, zato je to povsem primerna domena za preučevanje. V prvem poglavju bomo na kratko orisali to igro in pokazali, zakaj menimo, da je vredno raziskovati v tej smeri. Kot najbolj uporabno znanje, ki ga lahko napovemo, smo identificirali gibanje cen posameznih kart na prostem trgu. Da bomo lahko uspešno in natančno napovedali, ali se bo ta dvignila ali spustila, pa bomo potrebovali več vhodnih atributov. Ta naloga se bo spustila v celoten proces, ki ga izkusimo ob tipičnem procesu strojnega učenja. Prva naloga

bo identifikacija potrebnih atributov, ki jih je mogoče zajeti z računalnikom. Ena od večjih težav, ki jo bomo rešili, je zajem teh podatkov na več načinov, naj bodo to skripte, ki vsak dan pregledajo spletne strani, ali pa spletni obrazci, kjer bomo podatke vnesli sami. V ta namen smo v sklopu te naloge razvili spletno stran MkmScraper, ki opravlja prav to nalogo. Ko bomo imeli zadostno količino podatkov, jih bomo pretvorili v obliko, razumljivo programskemu paketu Weka. Zbrane podatke bomo uvozili v programski paket Weka ter pognali algoritem, ki smo si ga izbrali na podlagi več strokovnih člankov, ki so preučevali sorodne tematike.

V nalogi smo prišli do ne preveč uspešnih rezultatov, saj je naš izdelani model pri kartah, ki nas najbolj zanimajo dosegel povprečno točnost 61% napovedovanja trenda cen, kar le za nekaj odstotkov izboljša privzeto natančnost (59% pri zelo redkih kartah), pri nekaterih najbolj priljubljenih kartah pa tudi do 69 %. Ugotovili smo, da vsekakor obstaja povezava med izbranimi atributi ter nihanjem cene kart, ni pa zelo velika. Pri redkih kartah je bil naš sistem neučinkovit, saj je gibanje pravilno napovedal v povprečno 52%, kar je manj kot privzeta točnost (57%) in ni dovolj za uporabo v praksi.

Poglavje 2

Definicija problema

2.1 Predstavitev domene

Igra s kartami Magic: The Gathering je trenutno največja igra tipa TCG. TCG je kratica za Trading Card Game, kar pomeni, da gre za igro s kartami, kjer karte, s katerimi igramo, niso vedno enake, kot je značilno za tarok ali remi, ampak so lahko poljubno kombinirane iz velikega množice obstoječih kart. Igralci si sestavijo kupček (ang. deck), ki ga praviloma sestavlja 60 kart, vzetih iz omejene množice dovoljenih kart, s katerim igrajo proti nasprotnikom, ki so storili isto.

Igre se lahko odvijajo doma z družino za kuhinjsko mizo, s prijatelji v gostilni, na turnirju v lokalni trgovini (slika 2.1) ali pa na odru velikih turnirjev, kjer se nagrade za zmagovalca dvignejo tudi do 40 tisoč dolarjev.

Karte igralci kupujejo na več načinov. Lahko jih pridobijo z nakupom predsestavljenega paketa določenih kart ali pa poizkusijo svojo srečo z nakupom razširitvnega paketa naključnih kart (ang. booster - razširitveni paketek z naključnimi kartami iz ene izdaje), kjer dobi veliko pogostih kart (ang. common), nekaj nepogostih kart (ang. uncommon) ter eno redko karto (ang. rare). Iz tega je razvidno, da so nekatere karte bolj pogoste od drugih, večinoma pa drži, da so redkejše karte, poleg tega, da jih je manj, tudi močnejše ter posledično bolj zaželene od drugih.

Slika 2.1: Slika z Magic turnirja v Ljubljani

Med kartami v kupčku je več ali manj sinergije; če igralec želi sestaviti močan kupček, s katerim bo lahko osvojil turnir, torej potrebuje točno določene karte. Kupovanje naključnih kart zato ni preveč učinkovita metoda, ker v tem primeru igralec zapravi veliko denarja, da se dokoplje do želenih kart. Kjer obstaja povpraševanje, se kmalu najde rešitev, kar nas privede do druge možnosti.

S porastom popularnosti igre Magic: The Gathering se je kmalu razvil tako imenovani sekundarni trg s kartami, kjer lahko igralci za določeno ceno kupijo točno določeno karto, ki jo potrebujejo za svoj kupček. Vsekakor vse karte nimajo enake cene, kot morda velja pri zbiranju sličic za albume, temveč je višina cene odvisna od številnih faktorjev. Prvi od teh je že ta, kako redka je karta - pogosta karta bo večinoma cenejša od redke. Cene se ravnajo po principu prostega trga - torej pod vplivom ponudbe in povpraševanja. Redkost karte pa vsekakor ni edini faktor, ki vpliva na ceno. V tej nalogi bomo te faktorje identificirali in ovrednotili, kako vplivajo na ceno posamezne

karte.

2.2 Predstavitev problema

Magic: The Gathering je igra, ki jo po svetu igra več kot 10 milijonov ljudi, kar pomeni, da se izmenja veliko kart in obrne veliko denarja. Čeprav karte izdaja samo matično podjetje Wizards of the Coast, od te industrije živi mnogo ljudi. Prevladujejo podjetja, ki organizirajo velike turnirje na svetovni ravni s povprečno udeležbo 2000 igralcev. Velik igralec v industriji pa so tudi prodajalci produkta ter preprodajalci kart. Prav v zadnji kategoriji menimo, da se obrne največ denarja, kar je razlog, da bomo to nalogo posvetili optimizaciji njihovega dela.

Preprodajalci kart igrajo zelo preprosto vlogo v celotnem ekosistemu. Na eni strani poskušajo čim ceneje priti do kart, kar lahko naredijo tako, da odpirajo razširitvene paketke kart ali pa kupujejo karte od igralcev, v nekaterih primerih tudi od drugih trgovcev. Večja je razlika med ceno pridobivanja kart ter ceno, za katero jo kasneje prodajo končnemu kupcu, ki je ponovno lahko kdo od igralcev ali drugi trgovec, večji bo dobiček poslovanja podjetja.

Ko trgovci nastavljajo ceno, se pogosto ozirajo na celoten trg in jo postavijo blizu trenutnega tržnega ravnovesja. Ker se cena ravna po principu prostega trga, jo pogosto regulirajo glede na povpraševanje in ponudbo. Če se posamezna karta dobro prodaja, ji sčasoma zvišajo ceno, če pa se v določenem obdobju ne proda nobena, pa ji ceno znižajo.

Kot smo opazili, cena niha več ali manj reaktivno - ceno trgovci spremenijo šele po nekem dogodku. Iz tega sledi, da so zaradi tega prodali določeno število kart pod ceno ravnovesja ali pa niso prodali določenega števila kart, ki bi jih, če bi pravočasno znižali ceno.

To težavo bi lahko enostavno odpravili, če bi imeli sistem, ki bi znal predvideti, kako se bo cena spreminjala v naslednji uri, dnevu ali mesecu. Tu nam na pomoč priskoči sodobna tehnologija, predvsem veda strojnega učenja (ang. Machine Learning). S pomočjo te vede bomo v tej nalogi

sestavili model, ki bo z najnovejšimi algoritmi ter premišljeno sestavljenimi vhodnimi podatki znal predvideti trend nihanja cen v prihodnosti.

2.3 Sorodna dela

Vsekakor ta naloga ne bo prva, ki se bo ukvarjala z uporabo strojnega učenja pri predvidevanju cen, saj je uporabna vrednost strojnega učenja na tem področju zelo velika. Pri raziskavah na tem področju prevladuje preučevanje nihanja cen vrednostnih papirjev na borzah, kajti tam so na voljo zelo natančni podatki o cenah ter mnogo virov kakovostnih informacij, s katerimi so strokovnjaki gradili svoj model. Nekaj raziskav je narejenih tudi na področju nihanja valut ter cen žlahtnih kovin, nafte in drugih surovin. Nekaj teh raziskav bomo obravnavali v razdelku 3.2.1.

Poglavje 3

Definicija modela

3.1 Identifikacija atributov

Da bi bile naše napovedi čim bolj točne, moramo v model vključiti vse pomembne atribute, ki jih je mogoče predvideti in zanje tudi poiskati ustrezne podatke. V nadaljevanju poglavja bomo predstavili atribute, ki jih bomo vključili v model, s kratko razlago, kako vplivajo na ceno kart.

V igri Magic je trenutno več kot 30.000 kart, od katerih se cena večini ne spreminja več, zato za namen te naloge ne bomo zajeli celotne množice kart. V naši nalogi se bomo omejili na določen format igre, imenovan Standard. V njem lahko igralci igrajo z omejenim naborom kart, ki zajema izdaje, ki niso starejše od dveh let. Ko je karta starejša od dveh let, ob izidu nove izdaje ta karta zapusti format. Standard je trenutno najbolj igran format, zaradi česar cene precej nihajo in so zanimive za raziskavo.

V drugem delu poglavja si bomo pobližje ogledali, kaj je strojno učenje in zakaj smo si za preučevanje problema izbrali prav to vedo. Pregledali bomo tudi nekaj strokovnih člankov, ki so že preučevali sorodno tematiko. V naslednjih razdelkih so opisani pomembni parametri (atributi) za napovedovanje.

3.1.1 Zgodovina cene

Zgodovina cene karte je eden od naših najbolj pomembnih podatkov, ki bo služil kot rezultat našega modela oziroma atribut, ki se ga bo naš model poizkušal naučiti izračunati iz ostalih atributov.

3.1.2 Pogostost karte

V igri Magic: The Gathering poznamo štiri tipe pogostosti kart - pogosta (common), nepogosta (uncommon), redka (rare) ter izjemno redka (mythic rare). Redkost kart se izraža v pogostosti pojavljanja v razširitvenih paketkih kart, kjer se večina kart odpre in pride na trg. Pričakujemo, da bodo bolj redke karte bolj odzivne na zunanje dogodke, in sistem temu primerno modelirali.

3.1.3 Frekvenca pojavljanja na velikih turnirjih

V današnji dobi interneta se informacije hitro širijo, zato tudi dobre strategije sestavljanja kupčkov ne ostanejo skrite. Igralci velikokrat poskušajo posnemati profesionalne igralce ter želijo sestaviti podobne kupčke. Če se karta pogosto pojavlja na turnirjih, bo povpraševanje zanjo močno poskočilo ter posledično dvignilo ceno. Če se karta neha igrati v najboljših kupčkih, ji cena začne padati. Turnirje bomo ovrednotili po njihovi pomembnosti, saj večji in medijsko bolj izpostavljeni turnirji spremembi cene dodajo večjo vrednost.

3.1.4 Datum izida izdaje

Karte v igri Magic izhajajo približno na tri mesece. Takrat v obstoječo množico kart vstopi med 200 in 350 novih kart. Na začetku, ko nihče nima novih kart, ponudba ponavadi zaostaja za povpraševanjem, kar se občuti v višjih cenah kart na začetku tega obdobja.

3.1.5 Datum izhoda iz formata

Ko karta doseže starost približno dve leti, zapusti format Standard. To pomeni, da se njena uporabnost močno zmanjša, kajti uporabna je samo še za večje in močnejše formate ter priložnostne igre, ki pa veliko manj vplivajo na ceno. Čim bližje bo karta temu datumu, strmejšo krivuljo padanja cene pričakujemo.

3.1.6 Pomembni turnirji v prihodnosti

Pred vsakim turnirjem, kjer morajo igralci kupiti karte za svoj kupček, se povpraševanje po teh kartah poveča. Te turnirje bomo ovrednotili po pomembnosti, saj menimo, da lokalni manjši turnirji na ceno vplivajo manj kot mednarodni turnirji z več tisoč igralci.

3.1.7 Cena ostalih kart v izdaji

Karte v igri Magic se ponavadi kupi v škatlah s po 36 razširitvenimi paketi. Te škatle imajo na trgu relativno stabilno ceno. Iz podatkov o verjetnosti odprtja kart ter njihovi ceni lahko izračunamo matematično pričakovano vrednost celotne škatle. Tako kot je cena škatle na trgu stabilna, je tudi matematični seštevek cen vseh kart v izdaji stabilen. Iz tega sledi, da podražitev cene ostalih kart v izdaji pomeni pocenitev preučevane karte. Za ta atribut pričakujemo nekaj zamika, kar bomo upoštevali tudi pri napovednem modelu.

3.1.8 Pojavljanje v člankih

Ker je igra Magic zelo popularna, se je sčasoma razvilo veliko spletnih portalov, kjer profesionalni igralci objavljajo članke s teorijo igre, strategijami, ocenami in napovedmi. Igralci njihove nasvete močno cenijo in jih upoštevajo. Predvidevamo, da se z objavo določene karte v članku zviša povpraševanje po karti in posledično zviša cena, ker avtorji večinoma pišejo o dobrih kartah

Instead of having 4 **Disdainful Stroke**, I had 2 **Become Immense** as a way to race Ramp before it lands **Ugin.** This is the same for Rally, except the matchup is so horrendous that 2 **Hallowed Moonlight** come in as well. Those also incidentally come in against tokens-based-**Secure the Wastes** decks.

Here is the list I played last weekend:

GW Megamorph

Lands
4 x Windswept Heath
4 x Wooded Foothills
4 x Flooded Strand
3 x Canopy Vista
6 x Forest
4 x Plains

Creatures
4 x Warden of the First Tree
4 x Avatar of the Resolute
2 x Hangarback Walker
4 x Deathmist Raptor
4 x Den Protector

Spells
4 x Gideon, Ally of Zendikar
4 x Dromoka's Command
3 x Silkwrap
2 x Stasis Snare

Sideboard
4 x Valorous Stance
3 x Surge of Righteousness
2 x Become Immense
2 x Hallowed Moonlight
2 x Mastery of the Unseen
1 x Silkwrap
1 x Quarantine Field

Slika 3.1: Odsek iz članka, ki viša popularnost posameznih kart.

in uspešnih strategijah. Primer članka s strani DailyMTG si lahko ogledate na sliki 3.1

3.1.9 Prisotnost sezone formata

V igralnem letu se izmenja več sezon, ki narekujejo, kateri format se igra na večini lokalnih in mednarodnih turnirjev. Tri sezone so formata Standard, ena pa nekega drugega formata. Ko je na vrsti sezona Standarda, gredo cene kart, ki jih preučujemo, navzgor, v neaktivnem obdobju pa cene počasi padajo.

3.1.10 Količina odprtih razširitvenih paketkov

Veliko turnirjev se igra v najbolj priljubljenem formatu Limited, kjer igralci igrajo s kartami, ki jih dobijo na tistem turnirju, ko odprejo razširitvene paketke. Te karte nato preidejo v obtok in višajo ponudbo kart. V tem formatu se vedno odpirajo samo najnovejše izdaje, zato to vpliva samo na karte novejših izdaj.

3.2 Strojno učenje

Ian Witten s kolegi strojno učenje (ang. Machine Learning) v svoji knjigi [12] definira kot proces luščenja prej nepoznanih informacij iz podatkov z gradnjo računalniških programov, ki avtomatsko presejajo podatkovne baze v iskanju zakonitosti in vzorcev.

Veda izhaja kot veja umetne inteligence ter statistike in se kot ideja pojavlja že v sredini dvajsetega stoletja, ko jo je Arthur Samuel [11] definiral kot študijsko področje, ki da računalnikom sposobnost učenja, brez da so za to eksplicitno programirani. Od takrat se ta veda hitro razvija, še večjo veljavo pa dobiva v zadnjih desetih letih, ko so se mediji za dolgoročno hrambo podatkov (trdi diski) močno pocenili in postaja trend shranjevanja vseh mogočih podatkov vedno bolj popularen. Za to vedo se pojavljajo tudi druga bolj popularna imena, kot sta podatkovno rudarjenje (ang. Data Mining) ter v zadnjem času Big Data.

Velika količina kakovostnih podatkov je vsekakor dobra, a nam sama po sebi ne koristi. Šele ko te podatke pretvorimo v človeku razumljive informacije, si lahko z njimi pomagamo pri razumevanju naše domene. Tukaj na vrsto pride veda strojnega učenja, ki nam s svojim širokim naborom algoritmov pomaga doseči cilj.

Uporabnost strojnega učenja na praktičnih primerih se veča iz dneva v dan. Naj naštejemo nekaj primerov:

- Prepoznava črk na slikah

Slika 3.2: Googlov samovozeči avtomobil

- Prepoznava gruč strank
- Samovozeči avtomobili (slika 3.2)
- Vremenske napovedi
- Napovedi trendov delnic na borzi
- Razumevanje besedila

V grobem delimo strojno učenje na tri podzvrsti: **Nadzorovano učenje** preučuje probleme, kjer podatki vsebujejo tako vhodne kot tudi izhodne vrednosti. Računalnik se v tem primeru na učnih podatkih uči sestaviti napovedni model, ki mu omogoča iz vhodnih podatkov priti do izhodnih podatkov. Z drugimi besedami, računalniku povemo, kaj točno se mora naučiti. **Nenadzorovano učenje** uporabljamo, ko imamo podatke, za katere ne vemo, kaj točno velja zanje. Nimamo nekega končnega rezultata, do katerega želimo priti, zato računalniku pustimo, da sam odkrije smiselne skupine v učnem prostoru. Podamo mu zgolj vhodne podatke. Z nenadzorovanim učenjem pogosto iščemo skrite vzorce v podatkih in združujemo podatke v gruče. **Spod-**

bujevano učenje uporabljamo, ko želimo računalnik naučiti določenega vedenja. Primeri so vožnja avtomobila, agentsko trgovanje na borzi ter igranje iger proti nasprotnikom.

V naši nalogi bomo obravnavali trende cen v igri Magic: The Gathering in se učili na podlagi velike količine vhodnih in izhodnih podatkov napovedati cene v prihodnosti, kjer izhodnega podatka ne bomo imeli. Zaradi tega razloga se bomo omejili na algoritme **nadzorovanega učenja**. Ker predpostavljamo, da se Magic karte večinoma ravnajo po načelih prostega trga, si bomo pomagali s podobnimi raziskavami, narejenimi na tem področju. Najbolj uporabljano je strojno učenje na preučevanju trga vrednostnih papirjev, tako da je največ raziskav narejenih prav na tem področju.

Moje znanje o strojnem učenju izvira predvsem iz predavanj dr. Andrewa Nga z naslovom Machine Learning, ki so brezplačno na voljo na spletnem portalu Coursera, ter iz knjig avtorjev Iana H. Wittna, Eiba Franka ter Marka A. Halla Data Mining: Practical Machine Learning Tools and Techniques [12]. Za poganjanje algoritmov bomo uporabili v knjigah opisano brezplačno programsko opremo Weka 3, ki izvira z Univerze v Waikatu na Novi Zelandiji, kjer so aktivni avtorji. V nadaljevanju si bomo ogledali in kratko opisali nekaj strokovnih člankov s podobno tematiko. S pomočjo člankov bomo poskusili najti skupne točke pri izbiri algoritma ter pridobili dodatne ideje za našo raziskavo.

3.2.1 Pregled člankov s podobno tematiko

Rohit Choudhry in Kumkum Garg v članku [3] predpostavljata, da trg z vrednostnimi papirji ni naključen, temveč predvidljiv, vsekakor pa izpostavita, da je zapleten za modeliranje in odvisen od številnih atributov, kot so splošne razmere v gospodarstvu, politični dogodki ter pričakovanja trgovcev. V članku predstavita svoj algoritem, ki združuje algoritem z metodo podpornih vektorjev (ang. Support Vector Machines) ter genetskega algoritma (Genetic Algorithm). Avtorja sta bolj kot predvidevanje cene v prihodnosti raziskovala verjetnost padca in rasti. Ob uporabi genetskega algoritma sta

pri rezultatih opazila za nekaj odstotnih točk izboljšano točnost napovedi glede na rezultate pridobljene z osnovnim modelom SVM (60 % namesto 56 % pri delnici Infosys).

Nesreen K. Ahmed et al. v članku [1] primerjajo različne algoritme strojnega učenja za preučevanje časovnih vrst (ang. Time Series). V raziskavi so zajeli deset algoritmov, med drugimi tudi nevronske mreže, več vrst regresije ter algoritem SVM, in jih preizkusili na podatkih časovnih vrst tekmovanja M3. Za ta tip problematike sta se najbolj izkazala večnivojski perceptron (ang. Multilayer Perceptron) in Gaussova regresija procesov (ang. Gaussian Process Regression). Poleg algoritmov so avtorji preučevali tudi različne metode predprocesiranja podatkov, ki so različno vplivale na učinkovitost algoritmov.

Robert P. Schumaker in Hsinchun Chen v članku [10] posežeta po analizi trga vrednostnih papirjev z uporabo novic in člankov, objavljenih na spletu. Raziskujeta gibanje cen S&P 500 podjetij in v raziskavi zajameta več kot devet tisoč člankov. Za algoritem izbereta metodo SVM s posebnimi modifikacijami. V nadaljevanju pa primerjata še različne metode prepoznave teksta, kot sta Proper Noun ter Bag of Words. V rezultatih poročata o 57,1% točnosti napovedi smeri spreminjanja cene.

S podobno tematiko se ukvarjata tudi Desh Peramunetilleke in Raymond K. Wong v članku [8], ki pa naslove novic uporabljata pri napovedovanju nihanja svetovnih valut. S preučevanjem z vsebino bogatih tekstov poizkušata napovedati smer gibanja valut v naslednji uri do treh ur.

L. J. Cao in Francis E. H. Tay raziskujeta področje napovedovanja finančnih časovnih vrst v svojem članku [2]. Na kratko zajameta uporabo metode SVM s prilagodljivimi parametri ter njeno uporabnost in učinkovitost tudi na področjih, za katere je bila prvotno mišljena (prepoznavanje vzorcev). Algoritem primerjata z večnivojskimi nevronskimi mrežami ter nevronskimi mrežami, osnovanimi na radialni jedrni funkciji. Na podatkih borze v Chicagu pokažeta, da njun predlagani algoritem premaga oba primerjalna algoritma v napovedovanju finančnih napovedi.

O uspešnosti uporabe metode podpornih vektorjev poročajo še številni drugi avtorji, ki so se ukvarjali s to tematiko. Do podobnih ugotovitev pri spremljanju učinkovitosti in točnosti so prišli tudi Kyoung-jae Kim [7], Paul D. Yoo et al. [13] in Bin Gui et al. [4].

3.2.2 Izbira algoritma

V tej nalogi ne bomo odkrivali novih algoritmov, bomo pa poizkusili najti algoritem, ki bo najbolj ustrezal našemu problemu. Izbira algoritma je vsekakor zelo pomembna, še bolj pomembni pa so podatki. To je opazil že Googlov znanstvenik Peter Norvig [5], ki pravi, da lahko pripišemo boljše rezultate računalniškega učenja predvsem boljšim in bolj obsežnim podatkom, ne pa vse boljšim algoritmom.

Skoraj v vseh člankih smo zasledili veliko učinkovitost **metode podpornih vektorjev** (SVM), zato smo se v tej nalogi odločili, da si ga pobližje ogledamo in preučimo. Člankom je skupno tudi to, da napovedujejo le vzpon in padec cen, ne pa natančne cene. Tudi mi smo se odločili za ta pristop k reševanju problema.

Metoda podpornih vektorjev je eden od algoritmov nadzorovanega učenja. Osnovna ideja algoritma je razdelitev prostora s hiperravninami, kjer se nahajajo naši podatki, s katerim računalnik učimo. Vsak od teh podatkov pripada enemu od dveh razredov. Primer razredov sta razred artiklov, ki se jim je cena dvignila, ter razred artiklov, ki jim je cena padla. Na tej hiperravnini želi algoritem potegniti navidezno mejo tako, da bo prazen prostor med našimi primeri čim obsežnejši. S pomočjo tega območja ob meji lahko kasneje klasificiramo tudi nove primere, za katere ne vemo, v kateri razred sodijo. Osnova aplikacija tega algoritma je bila prepoznava vzorcev, trenutno pa se v industriji uporablja še za številne druge namene.

Poglavje 4

Implementacija

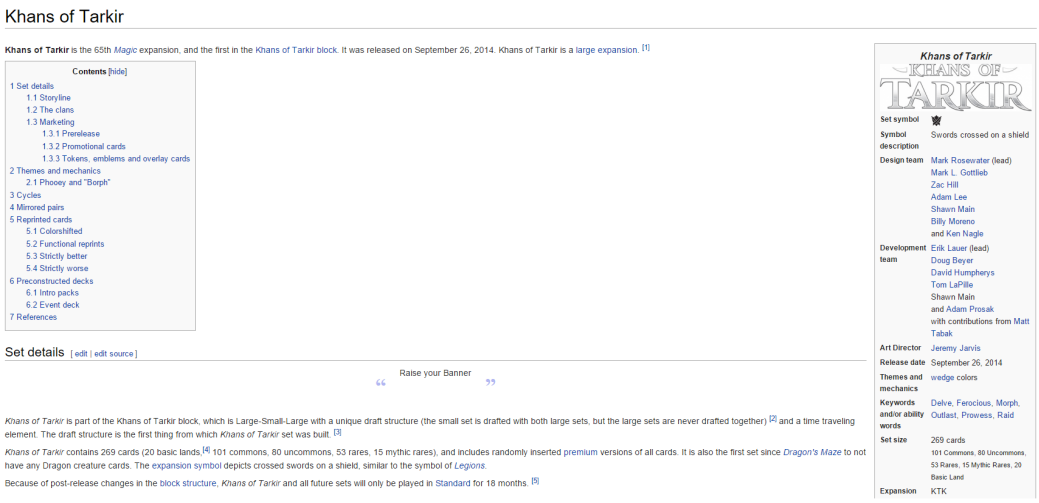
Sedaj ko smo definirali, kakšni bodo vhodni podatki za naš sistem predvidevanja cen, se lahko lotimo praktičnega dela. Naša prva naloga bo pridobivanje potrebnih podatkov iz več spletnih virov. Nadaljevali bomo s tem, kako te podatke pretvoriti v obliko, primerno za algoritem strojnega učenja.

4.1 Pridobivanje podatkov

Čeprav so podatki javno dostopni in dosegljivi, do njih ni enostavno dostopati v kakršnikoli primerni obliki. Velik del našega praktičnega dela bo tako obsegalo strganje podatkov s spleta ter beleženje v lokalno shrambo podatkov.

Podatki, ki jih bomo zajeli v tej nalogi, bodo omejeni na obdobje med septembrom 2015 ter decembrom 2015, in sicer predvsem zaradi pomanjkanja podatkov o zgodovini cen ter zapletenosti in zamudnosti pobiranja ostalih podatkov. V omenjenem obdobju bomo zajeli vse najpomembnejše dogodke, kot je rotacija formata, izid nove izdaje ter prisotnost velikih turnirjev.

V okviru te diplomske naloge smo sprogramirali spletno aplikacijo MkmScraper, ki smo jo prosto dostopno objavili tudi na portalu GitHub na naslovu `https://github.com/neyko5/mkmscraper`. Spletna aplikacija je napisana v jeziku PHP s pomočjo programskega ogrodja (frameworka) Laravel



Slika 4.1: Wiki stran s podatki o kartah

5. V ozadju teče podatkovna baza MySQL, za strganje podatkov pa skrbi knjižnica Goutte. Izbira tehnologij večinoma sloni na prejšnjem poznavanju in izkušnjah z njimi ter ni nujno potrebna za doseganje cilja.

4.1.1 Informacije o izdajah

V naši nalogi bomo obravnavali okrog deset različnih izdaj kart Magic, tako da smo se odločili za preprost obrazec na naši spletni aplikaciji, kjer je mogoče dodajanje, urejanje in brisanje izdaj. Za vir podatkov smo uporabili MtgSalvation Wiki (slika 4.1), ker je najbolj pregleden in temeljit. O izdajah bomo zbirali naslednje podatke:

- Ime izdaje
- Unikatno tričrkovno kratico
- Število kart v izdaji
- Število kart za posamezno pogostost
- Datum izida
- Datum rotacije iz formata Standard

4.1.2 Seznam kart

Kart, ki jih bomo preučevali, je več kot 2000, tako da ročno vnašanje ni najboljša možnost. Seznam kart je na voljo na več spletnih portalih, vendar so večinoma zelo okorni in neprimerni za branje z računalnikom.

Trgovanje z Magic kartami na spletu se večinoma odvija na dva načina. Prvi način je preko klasičnih trgovin, ki same postavljajo nakupno in prodajno ceno. Drugi način, ki je nam bolj zanimiv, pa je prek spletnih tržnic, podobnih portalu eBay.com, kjer kupci in prodajalci vzajemno upravljajo s cenami artiklov.

Za lažje branje smo se zatekli k pregledu ponujenih API-jev, ki so na voljo večinoma pri večjih tržnicah s kartami. Prvi, ki smo ga preizkusili, je bil delno zasebni API portala TCG Player, ki je eden večjih na področju proste tržnice s kartami tipa TCG. Na žalost nam avtorji niso dovolili dostopa, saj jim naša aplikacija v trenutni obliki ne bi prinašala prihodka.

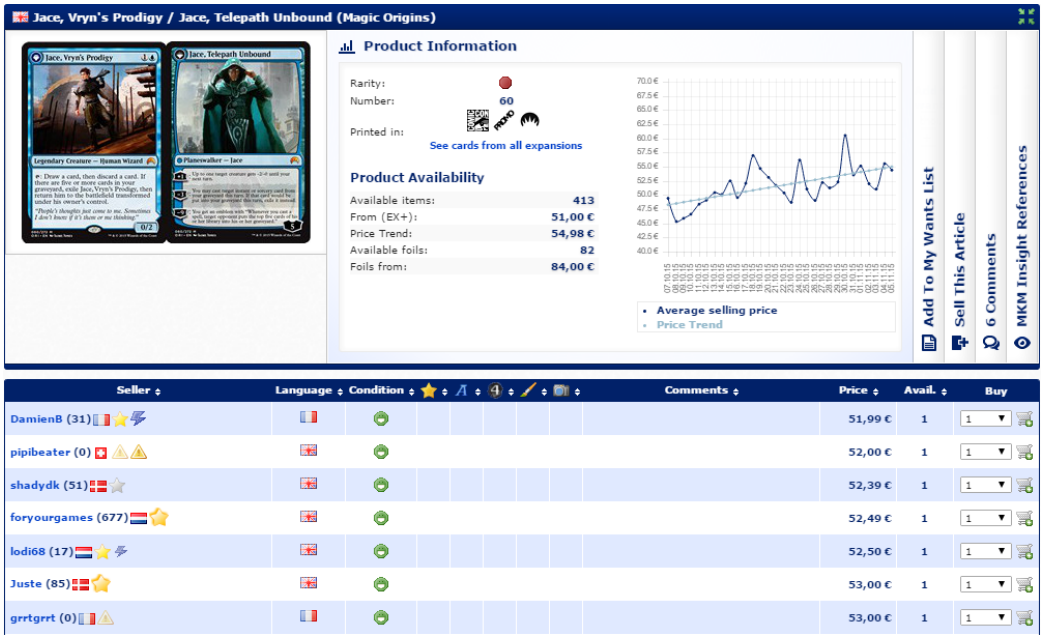
Srečo smo kasneje poizkusili pri največji evropski tržnici Magic Card Market. Njihov API je prosto dostopen, vendar zelo slabo dokumentiran. Ponuja pregled različnih iger (med drugim tudi Magic), izdaj in kart. Omogoča pregled trenutnih cen kart ter upravljanje inventarja posameznega uporabnika. Iz imena te tržnice izhaja tudi ime naše aplikacije - MkmScraper.

S pomočjo API-ja smo za vsako izdajo v našo podatkovno bazo shranili zapis za vsako posamezno karto s pripadajočimi podatki:

- Identifikacijska MKM številka
- Ime
- Izdaja
- Pogostost

4.1.3 Cene kart

Pridobivanje cen kart je vsekakor ena najpomembnejših nalog, ki jih moramo opraviti, saj predstavlja osnovo, na kateri smo zgradili naš sistem. Nobena od spletnih strani, ki se ukvarja s prodajo kart, ne ponuja dostopa do zgodovine



Slika 4.2: Primer prikaza karte na tržnici MKM

cen, tako da smo si morali zgodovino kart ustvariti sami. Odločili smo se, da bo najbolj primerno, da spremljamo ceno kart dnevno, saj je dan interval, ko se cena lahko dejansko spremeni in ni samo šum.

Cene smo spremljali na tržnici Magic Card Market, od koder smo pobirali že podatke o kartah. Čeprav njihov API ponuja tudi dostop do podatkov o cenah, njihov strežnik ne dovoli več kot tisoč dostopov v kratkem časovnem obdobju - v tem primeru blokira API-ključ. Zaradi te omejitve smo se morali zateči k manj primerni metodi luščenja podatkov iz HTML dokumentov.

Sprogramirali smo funkcijo, ki vsakih par minut poizkusi pridobiti dnevno ceno naključnih kart. Če bi želeli naenkrat pridobiti ceno vseh kart, bi preobremenili strežnik in nazaj dobivali napake. Za vsako karto obstaja unikatna povezava, sestavljena iz imena izdaje ter imena karte. Na primer za karto Mantis Rider iz izdaje Khans of Tarkir je povezava sledeča: */Products/Singles/Khans+of+Tarkir/Mantis+Rider*.

S pomočjo crawlerja Goutte naložimo stran posamezne karte ter poiščemo HTML značke, ki vsebujejo ceno, ter jo shranimo v našo podatkovno bazo.

V samem HTML dokumentu lahko enostavno dostopamo do naslednjih podatkov:

- Število kart na voljo (Available items)
- Najnižja cena za dobro ohranjeno karto (Price EX+)
- Cenovni trend (Trend price)
- Podatki za premijske karte (Foil price)

Premijske karte so bolj redke in cenjene različice kart, katerih cena se obnaša enako kot običajnim in smo jih zaradi pomanjkanja doprinosa k raziskavi zaenkrat preskočili.

Cenovni trend je nekakšna poenostavljena različica naše naloge. Cenovni trend se na portalu MKM računa tako, da se poišče enostavna linearna funkcija za regresijo podatkov prodajne cene zadnjih 30 dni, in se z njeno pomočjo izračuna ceno, ki se pričakuje na današnji dan. Regresijska funkcija, s katero se napoveduje trend cene je osnovna funkcija linearne regresije:

$$y = a * x + b \tag{4.1}$$

Cenovni trend je relativno nenatančna napoved cene, saj se, kot lahko vidimo na sliki 4.2, skoraj vedno zelo razlikuje od dejanske cene tisti dan. Eden od naših ciljev je, da premagamo napovedi linearne regresije oz. napovedi cenovnega trenda, ki ga modelira stran MagicCardMarket.

Naš najbolj pomemben podatek bo najnižja cena karte, saj je ena od boljših ocen trenutne tržne cene karte. MKM pri tem izračunu upošteva samo karte, ki so dobro ohranjene, in ignorira vse karte, ki so označene slabše kot *Excellent*.

Še boljši pokazatelj cene karte v določenem dnevu pa je povprečna vrednost dejansko prodanih kart v določenem dnevu. Tega podatka v HTML značkah sicer nimamo, se pa podatek pojavi na grafu zgodovine cen. Če skrbno preučimo, kako se graf sestavlja v kodi JavaScripta, lahko iz tam izluščimo podatke, ki jih potrebujemo. Za ta namen smo sprogramirali posebno funkcijo, ki naredi prav to.

Število kart, ki so na voljo, je eden od podatkov, ki nam pri modeliranju ne bo veliko koristil, saj se spreminja sočasno s ceno. Nihanje števila kart na voljo nam pri napovedi nihanja cene za naslednji dan tako ne pomaga pri večji natančnosti. Da bi lažje razumeli to razmerje lahko podamo primer iz druge domene. Relacija teh dveh parametrov je podobna kot relacija med povprečno hitrostjo smučarja ter čas, ki ga je potreboval za vožnjo po progi. Če imamo enega od teh podatkov, lahko iz njega izračunamo drugega. Vsekakor pa bo podatek o številu kart ob vizualizaciji podatkov na grafih lahko pokazatelj določenih trendov v ceni. Na podlagi preučevanja teh grafov bomo lahko v prihodnje bolje opazovali, kako sta cena in število artiklov povezana, in se naučili, kako prepoznati naravno in umetno manipulacijo cene. Če se kdaj odločimo za implementacijo sistema za prepoznavanje anomalij, bo to eden ključnih podatkov.

4.1.4 Frekvenca igranja posameznih kart

Da bi lahko prišli do tega podatka, smo morali poseči po zgodovinskih podatkih s turnirjev, natančneje po seznamu kupčkov, ki so se igrali na prejšnjih turnirjih (ang. decklists). Te lahko najdemo na številnih straneh, od uradnih strani turnirjev, kot sta Magic oddelek na strani Wizards of the Coast in StarCityGames.com, kot tudi na straneh, ki se ukvarjajo s strategijo in poročili s turnirjev, denimo MTG Goldfish ter MTG Top8.

Zanima nas, kdaj se je katera karta igrala, kakšno uvrstitev je dosegel kupček s to karto, kako pomemben je bil turnir ter v kakšni količini se je karta pojavljala (igralci lahko igrajo v kupčku do 4 enake karte). To smo ponovno storili s pomočjo webcrawlerja, ki nam v zameno za povezavo do posameznega turnirja v bazo shrani potrebne informacije, ki jih potrebujemo. Nekatere podatke, kot sta datum turnirja in njegov rang, smo vnesli sami. Za določanja ranga smo uporabili štiri vnaprej določene kategorije, ki so določene predvsem na podlagi gledanosti in obiskanosti turnirja. Turnirji vrste Pro Tour, ki imajo največjo veljavo, bodo ocenjeni z rangom 1, manjši lokalni turnirji pa z rangom 4.

Slika 4.3: Enostaven vmesnik za vnos kupčkov

Ker vsaka od omenjenih spletnih strani uporablja svojo HTML-kodo in značke, smo potrebovali štiri različne algoritme, da smo lahko shranili podatke iz čim več virov. Postopki se malenkostno razlikujejo, osnovni koraki pa so naslednji:

1. Naloži spletno stran vhodne povezave
2. V podatkovni bazi ustvari turnir z vsemi metapodatki
3. Poišči značko, ki vsebuje seznam kupčkov
4. Trenutno mesto nastavi na 1
5. Za vsak kupček stori naslednje:
   (a) Naloži spletno stran s kupčkom
   (b) Poišči značko s količino karte
   (c) Poišči značko z imenom karte
   (d) V DB poišči karto z istim imenom
   (e) Shrani zapis v DB
   (f) Povečaj mesto za 1

Za vnos smo pripravili enostaven vmesnik (slika 4.3), ki od uporabnika zahteva le povezavo do turnirja ter osnovne metapodatke.

4.1.5 Frekvenca pojavljanja v člankih

Članki so v skupnosti igralcev Magica zelo priljubljen način pridobivanja znanja o strategiji in sledenju trendov, zato jih ne smemo zanemariti. Za luščenje člankov smo si izbrali štiri najbolj popularne strani:

- StarCityGames.com
- Channel Fireball
- BlackBorder.com
- TCGPlayer.com

Vmesnik je zelo podoben kot pri vnašanju kupčkov, algoritem pa je še enostavnejši. Pri vsakem članku bomo ročno vnesli še datum objave. Pri vsakem članku bomo na spletni strani ugotovili, kje se nahaja naslov, ter izluščili vsebino, vse skupaj pa bomo shranili v našo podatkovno bazo. Sproti bomo vodili seznam vnesenih člankov (slika 4.4) in ga ažurno posodabljali.

4.1.6 Podatki o turnirjih

Turnirje bomo vnašali ročno, saj jih bo v naši nalogi okoli sto in ne obstaja preprostejša alternativa. Vnašali jih bomo preko preprostega vmesnika in zahtevali naslednje podatke:

- Ime turnirja
- Datum
- Rang

Entered Articles

ID	Title	Date	Popularity	Publisher
7	Daily Digest: Hot And Fresh Out The Kitchen	2015-11-04	1	StarCityGames
11	Standard Eldrazi and Aristocrats	2015-10-27	3	BlackBorder
12	Dark Jeskai at Pro Tour Battle for Zendikar	2015-11-04	1	ChannelFireball
13	Trees, Esper, and Scales	2015-11-03	2	TCG Player

Slika 4.4: Seznam vnesenih člankov

4.1.7 Podatki o sezonah in obdobjih

Prav tako bomo podatke o sezonah, kjer se igra format Standard, v glavnem vnašali ročno. Enako bomo storili z obdobji in njihovimi relativnimi količinami odprtih razširitvenih paketkov.

4.2 Priprava podatkov

Ko imamo v podatkovni bazi veliko količino podatkov, je naša zadnja naloga pred poganjanjem algoritmov učenja pretvorba podatkov v obliko, ki je primerna za vnos v programski paket Weka. Osnovna naloga bo normalizacija vseh atributov v območje med 0 in 1, kot predlaga Chih-Wei Hsu [6]. Normalizacija precej pomaga, da algoritmi ne izpostavijo atributov, ki imajo številčno najvišje vrednosti. Pri normalizaciji smo uporabili formulo:

$$x' = \frac{x - min(x)}{max(x) - min(x)}$$

Opisali bomo nekaj spremenljivk in odločitve, ki smo jih sprejeli zanje. Izhodno vrednost (skalar) bomo označili kot $\mathbf{y}$, vhodno vrednost (vektor) X, njegove posamezne komponente pa kot x_0,x_1,..,x_n.

4.2.1 Izhodna vrednost

Za izhodno spremenljivko $\mathbf{y}$ bomo izbrali povprečno prodajno ceno. Ker je premik cene na dnevni ravni zelo naključen in nepredvidljiv, smo se odločili opazovati ceno za obdobje enega tedna. Vzeli smo povprečno prodajno ceno na trenutni datum ter ji odšteli ceno na datum pred enim tednom. Pozitivni premiki so zato predznačeni pozitivno, negativni pa negativno. Ker bomo pri algoritmu pozorni samo na dviganje ali nižanje cene, bomo pozitivne premike označili kot razred 1, negativne pa bomo označili z -1 ter s tem izoblikovali dva izhodna razreda.

$$y' = sign(y_{today} - y_{lastweek})$$

4.2.2 Število igranih kart na turnirjih

Ker se turnirji odvijajo večinoma samo ob koncih tedna, se bodo podatki posodabljali tedensko. Ker preučujemo nihanje cen, nam več kot golo število igranih kart pomeni razlika glede na predhodni teden. Pridobiti želimo delež

naše karte glede na vse karte, igrane v preučevanjem obdobju, v račun pa želimo vključi doseženo mesto (številka med 1 in 8, ki jo izračunamo tako, da od 9 odštejemo doseženo mesto), število kopij karte ter rang turnirja, ki je naveden v spodnjem seznamu.

- 4 - Pro Tour
- 3 - Grand Prix
- 2 - Veliki turnirji
- 1 - Lokalni turnirji

Vrednost x_0 bomo pridobili s pomočjo spodnje funkcije in bo predstavljala razliko med prisotnostjo karte trenutni teden in prisotnostjo iste karte en teden prej. Ker domnevamo, da ima rang turnirja veliko večji vpliv kot karkoli drugega, bomo to vrednost kvadrirali.

Čeprav bomo parameter kasneje normirali, bo kvadriranje vseeno pravilno utežilo pomembnost ranga turnirja. Eden od primerov je primerjanje naslednjih dveh kart: Karta A je dosegla 1. mesto (8 točk) na lokalnem turnirju (1 točka), medtem ko je karta B dosegla 3. mesto (6 točk) na turnirju Pro Tour (4 točke). V primeru brez kvadriranja ranga bomo dobili vrednost 8 za karto A ter vrednost 24 za karto B. V primeru kvadriranja ranga pridemo do vrednosti 8 za karto A ter vrednosti 96 za karto B. Tudi ko vrednosti normaliziramo, je razlika med kartama A in B veliko večja v primeru kvadriranja ranga, kar pa je želen rezultat.

$$x0 = \frac{CardThisWeek * Rank^2 * Place}{AllThisWeek * Rank^2 * Place} - \frac{CardLastWeek * Rank^2 * Place}{AllLastWeek * Rank^2 * Place}$$

Enako bomo storili za x_1, le s to razliko, da bomo gledali podatke en teden v preteklosti.

4.2.3 Pojavljanje v člankih

Podobno kot z deležem igralnega časa bomo storili tudi s pojavljanjem v člankih. V tej nalogi ne bomo uporabili naprednih algoritmov razumevanja

tekstov, temveč bomo le prešteli članke, ki vsebujejo ime naše karte. Ponovno bomo gledali za obdobje enega tedna. Ker menimo, da na gibanje cen vpliva tako število objavljenih člankov kot tudi razlika med dejanskim in predhodnim tednom, bomo v naš model vključili oba atributa, poimenovana x_2 ter x_3. Prav tako bomo šli pri obeh atributih en teden v preteklost in atributa označili kot x_4 in x_5.

4.2.4 Sezona formata

Ker pri sezoni obstaja samo možnost, da sezona je ali pa je ni, bomo ta atribut definirali binarno ter ga označili z x_6.

4.2.5 Oddaljenost od izida in rotacije

Oddaljenost bomo definirali kot število dni od ročno vnešenega datuma. Oddaljenost od izida bomo označili kot x_7 ter oddaljenost od rotacije kot x_8. Ker je bližina teh skrajnih datumov pomembna največ mesec dni, bomo vrednosti preoblikovali s funkcijo, ki strmo narašča, ko se bliža ničli; eksponentno pada, ko se od nje odmika; ter se približa vrednosti 0, ko x prečka vrednost 30:

$$x' = \frac{1}{(e^x)^{0.2}}$$

4.2.6 Bližina turnirjev

S spremenljivko x_9 bomo označili število turnirjev v formatu Standard, ki povzročijo dvig cen kart. Vsak turnir bomo pomnožili s svojo ročno vneseno spremenljivko, ki bo osnovana na štirih pomembnostnih razredih turnirjev, navedenih zgoraj.

4.2.7 Trend ostalih cen v setu

Za vsako ostalo karto v izdaji bomo primerjali, koliko se je cena spremenila v zadnjem tednu, te vrednosti sešteli ter rezultat zapisali v x_{10}.

4.2.8 Količina odprtih razširitvenih paketkov

Količino bomo zajemali kot število odprtih paketkov na tipičnem Limited turnirju, na katerem vsak igralec odpre 6 paketkov. Vrednost bo vsebovala vrednosti med 0 in 6 ter nosila oznako x_{11}.

4.2.9 Tabela komponent

Za lažji pregled prilagamo tabelo vseh komponent, ki jih bomo uporabili v nalogi.

Oznaka	Opis
y	Povprečna prodajna cena
x_0	Razlika v igranosti karte ta teden
x_1	Razlika v igranosti karte prejšnji teden
x_2	Količina člankov ta teden
x_3	Razlika v količini člankov ta teden
x_4	Količina člankov prejšnji teden
x_5	Razlika v količini člankov prejšnji teden
x_6	Prisotnost sezone formata
x_7	Oddaljenost od izida
x_8	Oddaljenost od rotacije
x_9	Število turnirjev v formatu Standard
x_{10}	Trend ostalih cen v setu
x_{11}	Količina odprtih razširitvenih paketkov

4.3 Izbira orodja

Za namen te naloge smo si izbrali prosto dostopno orodje Weka 3. Vsi najbolj pogosti algoritmi, ki se uporabljajo v industriji, so kvalitetno napisani ter močno optimizirani, zato smo mnenja, da bi pisanje svojega algoritma presegalo namen tega dela. Weka že vsebuje algoritem SVM, ki ga bomo uporabili, in je enostavna za uporabo. Orodje za vnos podatkov uporablja format .arff, omogoča pa tudi neposreden dostop do podatkovne baze.

Poglavje 5

Rezultati

5.1 Izbira testnih podatkov

Podatke smo zbirali od 1. septembra 2015 do 10. decembra 2015. Izvozili smo jih za vsak dan posebej ter jih s pomočjo skripte, napisane v ogrodju Laravel, preoblikovali v format, ki je razumljiv Weki. Za posamezno karto smo imeli približno 100 učnih primerov (ang. training data), po en primer za vsak dan opazovanja. V podatkovno bazo je bilo vnesenih **310 člankov** ter seznami kupčkov s **180 turnirjev**.

V okviru naše naloge smo se za pridobivanje rezultatov odločili za preiskovanje kart izdaje Magic Origins. Izdaja je izšla približno en mesec pred začetkom pridobivanja podatkov ter je bila na splošno najbolj zanimiva za preučevanje, predvsem zaradi velike količine podatkov, ki smo jih lahko zbrali. Omejili smo se na karte z redkostjo rare ter mythic rare, saj so to v veliki večini primerov karte, s katerimi se največ trguje in bi jih bilo smiselno preučevati.

Za namen raziskave smo preučili 12 naključno izbranih kart z redkostjo rare ter vseh 15 kart z redkostjo mythic rare. Za vsako od teh kart smo posebej pognali algoritem. Na koncu smo v eno datoteko izvozili še vse mythic rare karte izdaje Magic Origins ter znova pognali algoritem.

5.2 Izbira parametrov

Za poganjanje smo na podlagi uspešnosti v več strokovnih člankih izbrali metodo podpornih vektorjev (SVM). Večkrat smo v različnih virih zasledili tudi algoritem sekvenčne minimalne optimizacije Johna Platta [9] (ang. Sequential minimal optimization - SMO), ki je modifikacija te metode in je že dolga leta standard za treniranje SVM-jev. Algoritem SMO olajša treniranje SVM-jev s tem, da zelo zahtevno reševanje problema kvadratičnega programiranja (ang. quadratic programming) razbije na več manjših problemov. Ti manjši problemi so rešljivi analitično, zaradi česar ni potrebe po potratnem numeričnem računanju. SMO zato porabi manj pomnilnika ter procesorske moči kot običajni SVM-ji, tako da je sposoben reševanja veliko večjih problemov. V tej nalogibomo ta algoritem uporabili za treniranje naših modelov, saj je popolnoma podprt v programskem paketu Weka.

Na primeru vseh mythic kart izdaje Magic Origins smo preizkusili več jedrnih funkcij (ang. kernel function) in njihovih parametrov. Najbolj se je obnesla polinomska funkcija tretje stopnje.

Za preverjanje točnosti smo uporabili metodo prečnega preverjanja s stopnjo k = 10 (cross-validation, folds 10). Ta metoda se uporablja pri testiranju točnosti modela za uporabo v praksi. Pri stopnji k = 10 vse učne podatke naključno razdelimo na 10 skupin. Devet skupin uporabimo za učenje modela, zadnjega pa uporabimo za preverjanje točnosti. Ta postopek ponovimo desetkrat s tem, da za testni del vedno izberemo drugo skupino.

5.3 Rezultati poganjanja algoritma

V spodnji tabeli so navedeni rezultati, ki smo jih pridobili s poganjanjem algoritma v programskem paketu Weka (tabela 5.3). V zgornjem delu tabele so navedene karte z redkostjo mythic rare, pod črto pa se nahajajo karte z redkostjo rare. Zadnji vnos pod drugo črto predstavlja skupen model za vse karte izdaje Magic Origins z redkostjo mythic rare. Za lažjo predstavo, kako nam Weka prikaže rezultate, prilagamo tudi zaslonski posnetek poročila

Weke, kjer smo trenirali model karte Jace, Vryn's Prodigy (slika 5.1).

```
Correctly Classified Instances          68               69.3878 %
Incorrectly Classified Instances        30               30.6122 %
=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 0.538     0.203      0.636     0.538      0.583      0.668    -1
                 0.797     0.462      0.723     0.797      0.758      0.668    1
Weighted Avg.    0.694     0.359      0.689     0.694      0.689      0.668

=== Confusion Matrix ===

  a  b   <-- classified as
 21 18 |  a = -1
 12 47 |  b = 1
```

Slika 5.1: Rezultati, vrnjeni s strani Weke za primer vseh zelo redkih kart. Vse točnosti so ocenjene z 10-kratnim prečnim preverjanjem

Karta	Točnost
Alhammarret, High Arbiter	55,102
Chandra's Ignition	47,959
Dark Petition	61,2245
Dwynen, Gilt-Leaf Daen	55,102
Flameshadow Conjuring	44,898
Hixus, Prison Warden	52,0408
Honored Hierarch	56,1224
Jace's Sanctum	45,9184
Kothophed, Soul Hoarder	44,898
Kytheon's Irregulars	56,1224
Languish	57,1429
Pia and Kiran Nalaar	53,0612
Tainted Remedy	50
Alhammarret's Archive	45,9184
Archangel of Tithes	57,1429
Avaricious Dragon	51,0204
Chandra, Fire of Kaladesh	62,2449
Day's Undoing	68,3673
Demonic Pact	59,1837
Disciple of the Ring	69,3878
Jace, Vryn's Prodigy	69,3878
Kytheon, Hero of Akros	58,1633
Liliana, Heretical Healer	69,3878
Nissa, Vastwood Seer	68,3673
Pyromancer's Goggles	54,0816
Starfield of Nyx	60,2041
The Great Aurora	65,3061
Woodland Bellower	59,1837
Magic Origins Mythics	62,2449

5.4 Povzetek rezultatov

Pri naših poizkusih smo preučevali dve dokaj različni kategoriji kart, ki sta se razlikovali po redkosti. Pred pričetkom testiranja smo pričakovali, da bodo napovedi bolj točne za dražje in redkejše karte, kar se je izkazalo za pravilno napoved. Modeli za karte z redkostjo rare so dosegali točnost med 44 ter 61 odstotki, povprečno pa so karte dosegale 52% točnost, kar je manj kot znaša privzeta točnost. Privzeta točnost je točnost, ki jo dobimo, če bi klasifikator vedno klasificiral v večinski razred. Ta je za redke karte znašala 57% (cena je padla v 57% primerov), kar pomeni, da bi brez učenja modela dosegli boljše rezultate v primeru da bi vedno predvidevali padec cene naslednji dan.

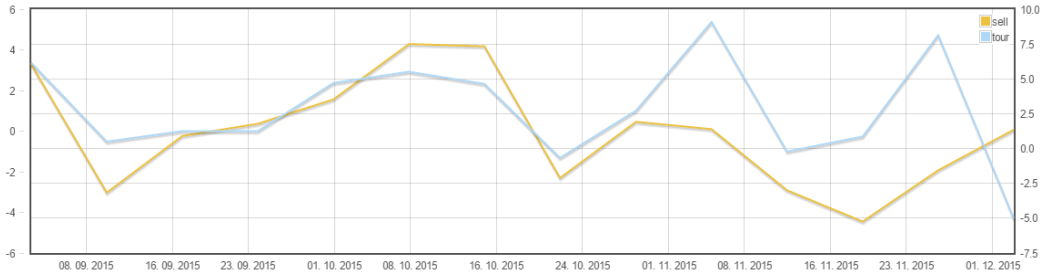
Modeli kart z redkostjo mythic rare so dosegali točnost med 45 ter 69 odstotki, povprečno pa 61 %. Privzeta točnost za mythic rare karte je znašala v povprečju 59% (padec cene), kar pomeni, da je naš model to točnost prekosil (čeprav ne za veliko) in bolje predvidel gibanje cene, kot bi jo predpostavljanje padca cene.

Najboljše rezultate so dosegale karte z visoko popularnostjo ter visoko ceno. Najslabše rezultate pa karte, ki jih nihče ne potrebuje in je njihova cena zelo nizka in se večinoma giblje naključno. Čeprav smo imeli za vse karte okoli 100 učnih primerov pa pri nepriljubljenih kartah nismo mogli pridobiti naših najbolj pomembnih atributov prisotnosti na turnirjih in člankih, saj se te karte tam sploh niso pojavljale. To je vsekakor vplivalo na točnost naših napovedi, saj je pri teh kartah ta občutno nižja. Rezultati, ki smo jih dosegli pri redkih kartah, imajo zato zelo malo praktične uporabnosti. Rezultati, ki smo jih pridobili s preučevanjem zelo redkih kart, predvsem pa tistih bolj priljubljenih, pa nam lahko v praksi z dovolj visoko točnostjo napovedi pripomorejo pri pridobivanju prednosti pred tekmeci v trgovanju s kartami Magic. Prodajalci kart v praksi večino dobička pridobijo s prodajo zelo vrednih in redkih kart, tako da so za praktično uporabo bolj pomembni rezultati za bolj redke in drage karte.

5.5 Nadaljnje odkrivanje znanj iz podatkov

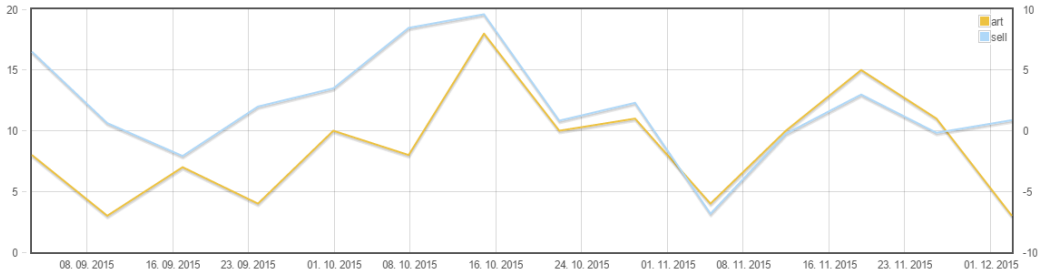
Ker pri sestavljanju modela nismo prišli do zelo visokih točnosti, smo želeli dobiti bolj poglobljeno sliko v podatke in njihovo odvisnost. Pričeli smo z risanjem grafov odvisnosti gibanja cene v primerjavi z enim od atributov z različnimi metodami računanja vrednosti. Primeri metod, ki smo jih uporabili, so:

- Krajšanje in daljšanje obdobij opazovanja
- Preklapljanje med absolutno vrednostjo in razliko med tedni
- Bolj ali manj ublaženo računanje razlik
- Premikanje časa opazovanja za n tednov v preteklost in prihodnost
- Uporaba drugih podatkov za ceno (najnižja vrednost namesto povprečna)
- Množenje več podatkov skupaj
- Uteževanje različnih podatkov



Slika 5.2: Graf odvisnosti gibanja cene od turnirske prisotnosti. Oznaka tour nam prikazuje prisotnost karte v zmagovalnih kupčkih na turnirjih, oznaka sell pa prodajno ceno.

Grafi so bili v večini primerov relativno nepovezani, po neštetih poskusih pa smo prišli do dveh grafov, ki sta se skoraj popolnoma ujemala. Prvi graf prikazuje prisotnost karte v zmagovalnih kupčkih na turnirjih in prodajno ceno (slika 5.2), drugi graf pa število omembe karte v člankih ter prodajno ceno (slika 5.3). To je vsekakor dobra novica za naš sistem, vendar smo pri tem naleteli na novo težavo. Oba grafa namreč podatke črpata tako, da za



Slika 5.3: Graf odvisnosti gibanja cene od prisotnosti v člankih. Oznaka art nam prikazuje prisotnost v člankih, oznaka sell pa prodajno ceno.

nastope na turnirjih ter prisotnost v člankih gledata za en teden (članki) oziroma dva tedna (turnirji) v prihodnost. Čeprav bi algoritmi za strojno učenje z veseljem uporabili te podatke, teh v praksi nimamo na voljo in si z njimi ne moramo pomagati. V tej nalogi tako nismo smeli uporabiti teh dveh parametrov v tej obliki, zato smo se odločili za različico parametrov, kjer časa ne prestavljamo v prihodnost. Ker je še vedno prihajalo do določene mere ujemanja med atributoma ter nihanjem cene se nam je zdela njihova vključitev smotrna.

Očitno je odvisnost cene in teh dveh atributov ravno nasprotna, kot smo predvidevali pri modeliranju sistema, kar nakazuje, da je glavni vzrok nihanja cen skrit nekje drugje in ga nismo uspeli odkriti, ali pa je ta skrit (najboljši igralci pred ostalimi odkrijejo, katere karte so dobre). Druga razlaga je, da je ta atribut slabo berljiv (video posnetki s turnirjev) ali pa ni berljiv za računalnik (ni dostopen na internetu). Ta ugotovitev pa nam lahko prav pride pri malce drugačnem problemu, in sicer pri napovedovanju enega od teh atributov glede na ceno. Tukaj se poraja vprašljivost praktične uporabnosti teh podatkov. Kako si lahko pomagamo s tem, da bo določena karta večkrat omenjena v članku? Verjetno ne kaj dosti. Morda bolj zanimiv podatek zna biti količina določene karte na turnirjih v prihodnosti, saj vsekakor prinaša strateško prednost igralcem, saj se lahko bolj natančno pripravijo na kupčke, ki bodo bolj pogosti.

Poglavje 6

Sklepne ugotovitve

Po analizi rezultatov smo ugotovili, da s pomočjo strojnega učenja nismo prišlo do preveč dobrih rezultatov, saj smo pri zelo redkih kartah komaj presegli privzeto točnost, pri redkih pa niti te ne. Razlogov za to je lahko več. Eden od zelo verjetnih je visoka zahtevnost domene in nepovezanost podatkov. Drug zelo verjeten razlog je nedostopnost ključnih podatkov na internetu kot so na primer rezultati priprav na turnirje najboljših ekip ter pojavljanje kart na popularnih video posnetkih.

Vsekakor model, ki smo ga ustvarili, ni popoln. Predvsem ga pesti nepopolno ujemanje najbolj pomembnih atributov, kot sta prisotnost na turnirjih in v člankih, ki sta, kot smo kasneje ugotovili, posledica in ne vzrok spreminjanja cen. V nadaljevanju vam bomo predstavili nekaj idej, v katero smer se lahko nadaljuje naša raziskava ter pripomore k večji praktični uporabnosti.

6.1 Naslednji koraki

V tej nalogi smo sestavili zelo osnoven model, ki je pokazal smiselnost implementacije in uporabe strojnega učenja v predvidevanju cen kart. Sistem vsekakor ni popoln, zato v prihodnje predlagamo naslednje korake k izboljšanju točnosti ter ideje za nadaljnji razvoj.

6.1.1 Več podatkov ter boljša kvaliteta

V tej nalogi smo zajemali podatke v obdobju nekaj mesecev, ko so se vsaj enkrat odvili vsi pomembni dogodki. Večja količina podatkov bi vsekakor izboljšala učinkovitost napovedi. Kvaliteto podatkov bi izboljšali tudi z luščenjem podatkov iz več zanesljivih virov ter primerjavo med njimi.

6.1.2 Večje število atributov

V nalogi smo večinoma črpali podatke iz lastnih izkušenj o domeni na podlagi desetletnega amaterskega udejstvovanja. Čeprav smo verjetno zajeli večino pomembnih faktorjev, pa zagotovo obstajajo ljudje, ki o domeni vedo več. S pogovorom z njimi bi lahko poiskali nove atribute in jih dodali v naš model ter ga s tem izboljšali.

6.1.3 Povezava s trgovino

Če bi naš sistem povezali s spletno, fizično ali pa celo simulirano trgovino s kartami, bi lahko iz prve roke spoznali, kako se naš sistem odziva na dejanskem trgu. Pravilna napoved ne pomeni nujno največjega mogočega dobička. Z učenjem na trgovanju in ocenjevanjem kvalitete algoritma na podlagi dobička bi lahko ustvarili nov nivo računalniškega učenja.

6.1.4 Implementacija agentov

Nov korak k avtomatizaciji trgovine z Magic kartami bi bila implementacija programskih agentov, ki bi lahko nadomestili trgovca. Agent bi lahko s pomočjo naših napovedi sam kupoval ter prodajal karte, prav tako pa bi lahko prihranil veliko časa in denarja, saj bi znal odločitve sprejemati sam. Če bi bil sistem napovedi kakovostno izdelan, pa bi lahko te odločitve celo prekašale trgovčeve.

Literatura

[1] Nesreen K. Ahmed, Amir F. Atiya, Neamat El Gayar, and Hisham El-shishiny. An empirical comparison of machine learning models for time series forecasting. *Journal of Econometric Reviews*, 29:594–621, August 2010.

[2] L.J. Cao and F.E.H. Tay. Support vector machine with adaptive parameters in financial time series forecasting. *Neural Networks, IEEE Transactions on*, 14(6):1506–1518, Nov 2003.

[3] Rohit Choudhry and Kumkum Garg. A hybrid machine learning system for stock market forecasting. *World Academy of Science, Engineering and Technology 39*, pages 201–202, 2008.

[4] Bin Gui, Xianghe Wei, Qiong Shen, Jingshan Qi, and Liqiang Guo. Financial time series forecasting using support vector machine. In *Computational Intelligence and Security (CIS), 2014 Tenth International Conference on*, pages 39–43, Nov 2014.

[5] Alon Halevy, Peter Norvig, and Fernando Pereira. The unreasonable effectiveness of data. *IEEE Intelligent Systems*, 24(2):8–12, March 2009.

[6] Chih W. Hsu, Chih chung Chang, and Chih jen Lin. A practical guide to support vector classification, 2010.

[7] Kyoung jae Kim. Financial time series forecasting using support vector machines. *Neurocomputing*, 55:307–319, March 2003.

[8] Desh Peramunetilleke and Raymond K. Wong. Currency exchange rate forecasting from news headlines. *Aust. Comput. Sci. Commun.*, 24(2):131–139, January 2002.

[9] John C. Platt. Sequential minimal optimization: A fast algorithm for training support vector machines. *Technical Report MSR-TR-98-14*.

[10] Robert P. Schumaker and Hsinchun Chen. Textual analysis of stock market prediction using breaking financial news: The azfin text system. *ACM Trans. Inf. Syst.*, 27(2):12:1–12:19, March 2009.

[11] Phil Simon. *Too Big to Ignore: The Business Case for Big Data*. Wiley, 1st edition, 2013.

[12] Ian H. Witten, Eibe Frank, and Mark A. Hall. *Data Mining: Practical Machine Learning Tools and Techniques*. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 3rd edition, 2011.

[13] P.D. Yoo, M.H. Kim, and T. Jan. Machine learning techniques and use of event information for stock market prediction: A survey and evaluation. In *Computational Intelligence for Modelling, Control and Automation, 2005 and International Conference on Intelligent Agents, Web Technologies and Internet Commerce, International Conference on*, volume 2, pages 835–841, Nov 2005.